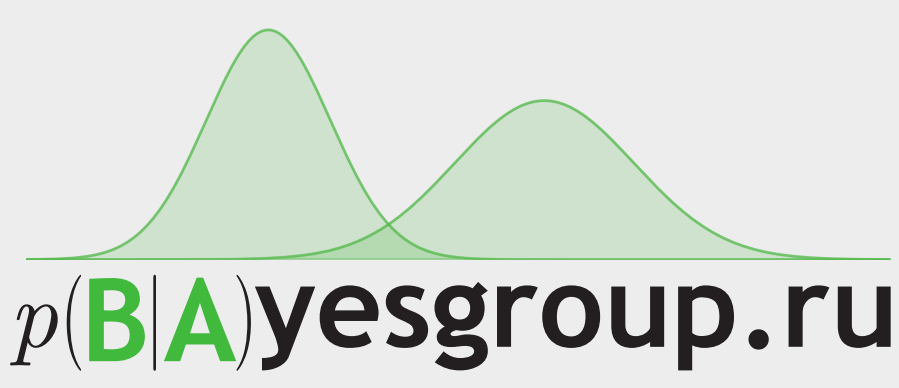


# Dropout-based Automatic Relevance Determination

Dmitry Molchanov  
dmitry.molchanov@skolkovotech.ru

Arseniy Ashuha  
ars.ashuha@phystech.edu

Dmitry Vetrov  
vetrovd@yandex.ru



## Main result

We propose a new way to train adaptive individual dropout rates for each model weight. Our approach provides Automatic Relevance Determination and enforces model sparsity.

## Motivation

- With Bayesian Machine Learning techniques we can construct sparse models like the Relevance Vector Machine
- Traditional Automatic Relevance Determination approach is hard to apply to deep neural networks
- Variational Dropout is an intriguing technique that allows to train individual dropout rates for each model weight, so we build upon this approach

## Variational Dropout model definition and training

- Approximate posterior distribution and log-uniform prior
$$q(w_i | \theta_i, \alpha_i) = \mathcal{N}(\theta_i, \alpha_i \theta_i^2) \quad p(\log |w_i|) \propto c$$
- Training: optimize the Variational Lower Bound w.r.t.  $\theta, \alpha$ 
$$\mathcal{L}(q) = \mathbb{E}_{q(w)} \log p(T | X, w) - D_{KL}(q(w) \| p_{prior}(w)) \rightarrow \max_{\theta, \alpha}$$
- When  $\alpha$  is fixed, VDO is equivalent to Gaussian Dropout:
$$w_i \sim q(w_i) \Leftrightarrow w_i = \theta_i \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(1, \alpha_i)$$
- Training procedure (Doubly Stochastic Variational Inference)  
**Require:**  $X$ : train objects;  $T$ : train labels;  $M$ : mini-batch size  
**Ensure:** optimal  $\theta, \alpha$ 
  - 1:  $\theta, \alpha \leftarrow$ : initial approximation;
  - 2: **repeat**
  - 3:  $X, T^M \leftarrow$  Random mini-batch of M datapoints
  - 4:  $\triangleright$  Sample gradient estimate using the Local Reparametrization Trick
  - 5:  $g \leftarrow \nabla_{\theta, \alpha} \hat{\mathcal{L}}^M(\theta, \alpha; (X, T)^M)$
  - 6:  $(\theta, \alpha) \leftarrow$  Update parameters using gradient-based method like Adam
  - 7: **until** convergence of parameters  $(\theta, \alpha)$
- Only  $\alpha \leq 1$  case is considered (authors report problems with large-variance gradients in case of large values of  $\alpha$ )

## Variational Dropout Automatic Relevance Determination

Main idea: driving dropout rates  $\alpha_j$  to infinity pushes  $\theta_j$  to 0 and enables Automatic Relevance Determination:

$$q(w_i | \theta_i, \alpha) = \mathcal{N}(w_i | \theta_i, \alpha_i \theta_i^2) = \delta(0)$$

To achieve this goal we offer

- Analytical analysis in case of linear regression
- A way to reduce gradient variance
- New approximation of KL divergence which is valid for all values of  $\alpha$
- Experiments that show that ARD effect persists in doubly stochastic setting

## Linear Regression

Likelihood:

$$p(t | X, w) = \mathcal{N}(t | Xw, \beta^{-1}I)$$

Objective:

$$\mathcal{L}(q) = \frac{N}{2} \log \beta - \frac{\beta}{2} \left[ \|X\theta - t\|^2 + \sum_d \kappa_d \theta_d^2 \right] - D_{KL}(\alpha)$$

Exact update for  $\theta$  given fixed  $\alpha$ :

$$\theta^* = (X^\top X - \text{diag}(\kappa))^{-1} X^\top t, \quad \kappa_d = \alpha_d \sum_{n=1}^N x_{nd}^2$$

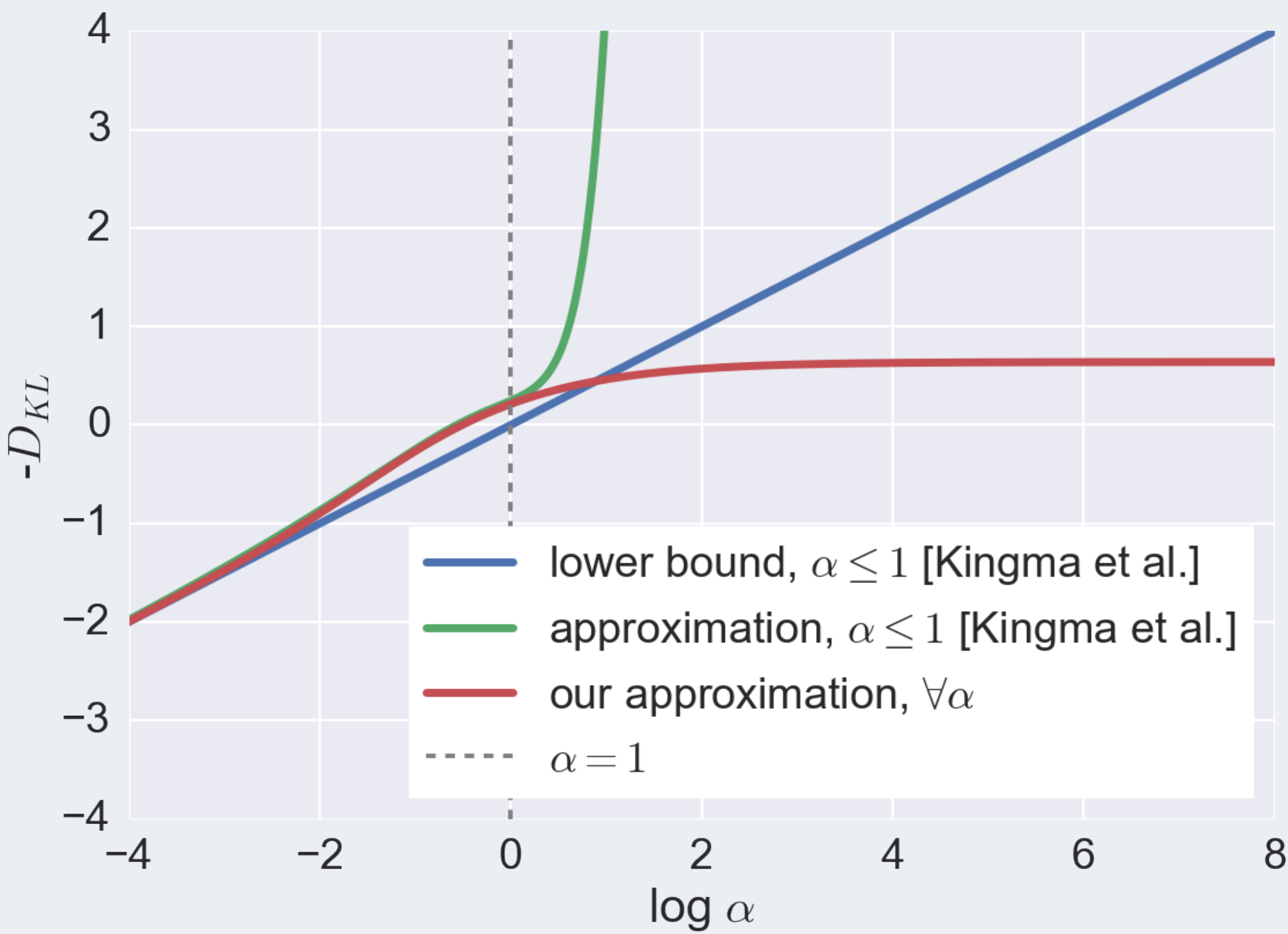
- Looks much like corresponding expression from the RVM regression:
$$w^* = (X^\top X - \text{diag}(\alpha))^{-1} X^\top t$$
- $\alpha_d \rightarrow +\infty \Rightarrow \mathcal{N}(\theta_d, \alpha_d \theta_d^2) \rightarrow \mathcal{N}(0, 0) = \delta(0)$

## New parametrization

- Original multiplicative noise yields noisy gradients:
$$w_i = \theta_i(1 + \sqrt{\alpha_i} \cdot \varepsilon_i), \quad \frac{\partial w_i}{\partial \theta_i} = 1 + \sqrt{\alpha_i} \cdot \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, 1)$$
- Proposed equivalent additive noise reduces variance of gradients:
$$w_i = \theta_i + \sigma_i \cdot \varepsilon_i, \quad \frac{\partial w_i}{\partial \theta_i} = 1, \quad \varepsilon_i \sim \mathcal{N}(0, 1)$$
- We now use  $q(w_i) = \mathcal{N}(w_i | \theta_i, \sigma_i^2)$  and can obtain  $\alpha_i$  from  $\alpha_i \theta_i^2 = \sigma_i^2$

## New Approximation for KL Divergence

- Original VDO paper offered several KL divergence approximation for  $\alpha \leq 1$
- To drive  $\alpha$ 's to infinity and achieve ARD effect we generalize approximation for all values of  $\alpha$  by sampling the divergence and solving regression task
$$-D_{KL}(q \| p) \approx 0.64 \sigma(1.5(1.3 + \log \alpha_i)) - 0.5 \log(1 + \alpha_i^{-1}) + c$$



- We also showed that  $\alpha \rightarrow +\infty \Rightarrow D_{KL}(\alpha) \rightarrow \text{const}$

## Automatic Relevance Determination in RVM vs ARD in VDO

The cause and nature of ARD in VDO is completely different from ARD in RVM

### ARD in RVM

- 1  $\alpha_i \rightarrow +\infty$
- 2  $p(w_i | \alpha_i, \text{Data}) = \delta(0)$
- 3  $\alpha_i$  are prior parameters
- 4 Fit prior distribution parameters to data
- 5 ARD by shrinking the prior to a delta function

### ARD in VDO

- 1  $\alpha_i \rightarrow +\infty$
- 2  $q(w_i | \theta_i, \alpha_i) = \delta(0)$
- 3  $\alpha_i$  are variational parameters
- 4 Prior distribution is fixed
- 5 ARD by introducing infinitely strong multiplicative noise

Our approach comes more naturally from the Bayesian framework, as the prior distribution should remain independent from the training data.

## Experiments

Model: VD-ARD multiclass logistic regression, trained with DSVI

Dataset	Accuracy			Sparsity		
	VD-ARD	L1-LR	RVM	VD-ARD	L1-LR	RVM
MNIST	<b>0.926</b>	0.919	N/A	<b>69.8%</b>	57.8%	N/A
DIGITS	0.948	<b>0.955</b>	0.945	<b>75.4%</b>	38.0%	74.6%
DIGITS + noise	0.930	<b>0.937</b>	0.846	<b>87.6%</b>	55.9%	85.7%

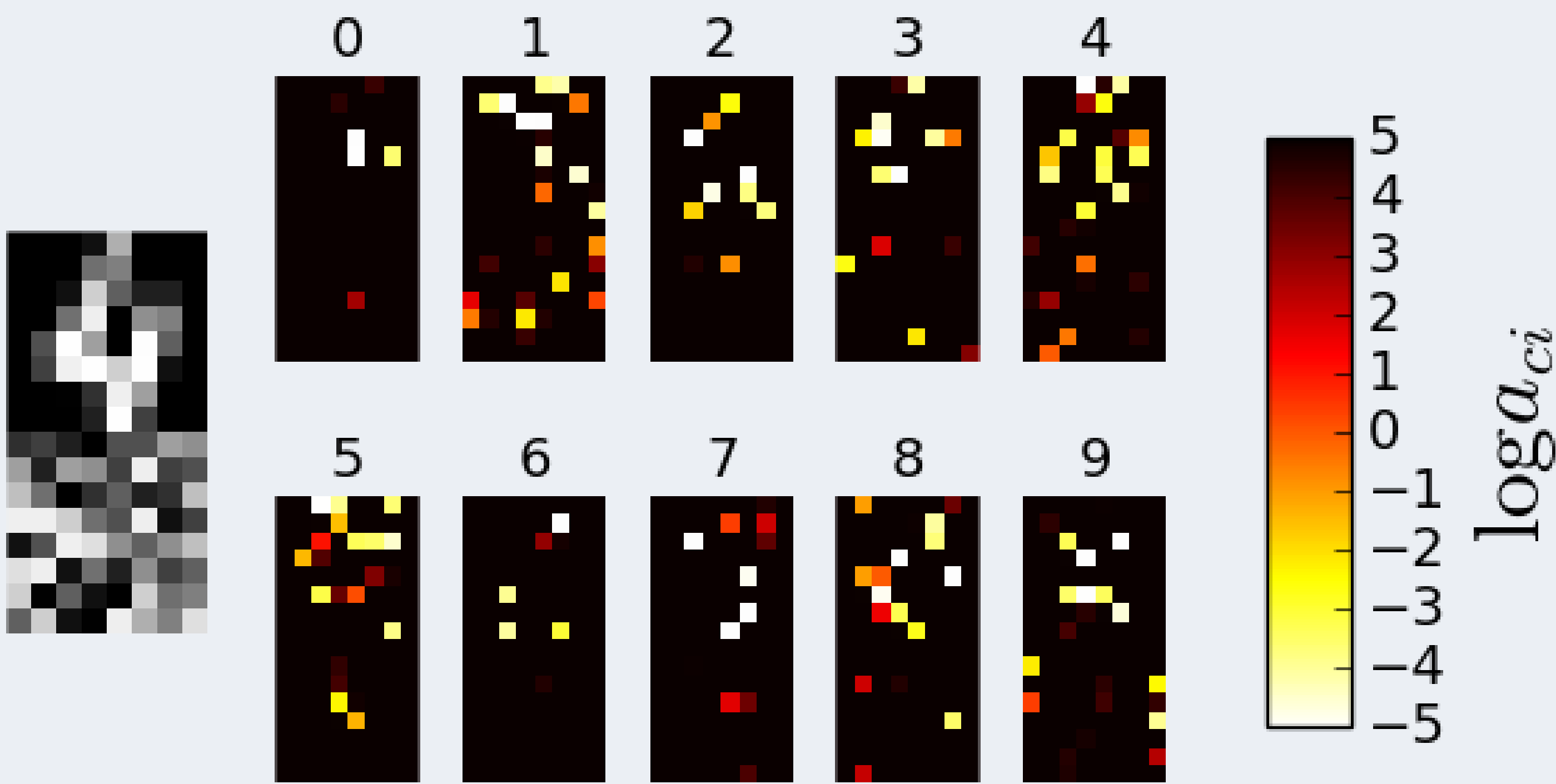


Figure: An example of an object with concatenated noise. Trained dropout rates.

## Discussion

- Variance of the gradients in VDO can be decreased by replacing multiplicative noise with equivalent additive noise
- Variational Dropout can be used to obtain ARD effect
- This effect still holds in doubly stochastic scenario (can be applied to DNNs!)
- Can't share  $\alpha$ 's and therefore can't enforce group sparsity